Selection of the Free Parameters in the Support Vector Method Using Bootstrap Resampling

Rojo-Álvarez, José Luis
Artés-Rodríguez, Antonio

Abstract—***The Support Vector Method is an efficient learning algorithm. However, the prior selection of its free parameters through either the cross-validation of the training set or the VC-theory bounds is problematic when dealing with low-sized data sets. We propose the boostrap resampling to yield robust estimates of the free parameters under these circumstances. The effectiveness of the method is evaluated using both synthetic and real data, showing this is an efficient framework when low number of data are available.***

Index terms*—*Support Vector Method, cross-validation, bootstrap, resampling, actual risk.

I. Background and Motivation

The Support Vector Method is a solid framework for general statistical learning problems [1,2,3]. At the heart of its advantages, it finds itself the intermediate feature space where the learning decision is arranged; its error function has no local minima, but rather an unique solution; and it can be formulated in non-linear problems through the use of different kernels .

However, the method makes no allowance for two open issues; the tuning of the *kernel parameter*, related to the generalization capabilities of the machine, and the *trade-off* between the margin (distance between classes) and the losses. With a large amount of data, conventional cross-validation can help find the parameters minimizing the estimated error probability on the validation subset. For low-size data sets, splitting the set into training and test subsets leads to losses in the generalization capabilities of the trained machine. A number of error bounds trying to base just upon the training set have been proposed, such as the VC-dimension or the leave-one-out [1]. These bounds become extremely weak for low-size data sets, and more, they cast some doubt when faced to non-separable data.

An estimation of the actual (i.e., whole, not only empirical) error probability upon the training set would allow to choice the free parameter minimizing this estimated error. The *bootstrap resampling* techniques [4] can be used to trace an estimate of the error probability having a reduced bias towards the empirical risk. Moreover, this procedure is robust when faced to low-size data sets, and it needs no assumptions on the statistical distribution from the data, so it can perfectly work under the same premises than the SVM requires.

The plan of the paper is as follows. Section II describes the SVM and the Vapnik's actual risk bound for a classification problem. We then present the formulation of the bootstrap estimate upon the error probability for this learning machine in Section III. Section IV contains a toy example showing the adequate performance of the procedure for linear and non-linear classifiers and for both free parameters. A real data problem, related to the automatic arrhythmia discrimination in implantable devices, is analysed in Section V. Finally, Section VI, conclusions are drawn.

II. The Actual Risk and The Support Vector Method

Be a set of observed, labelled data:

$$\mathbf{V} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_l, y_l)\}, \quad (1)$$

where $\mathbf{x}_i \in \mathbf{R}^n$ and $y_i \in \{+1,-1\}$. Be a non-linear transformation $\phi(\mathbf{x}_i)$ to a usually unknown, higher dimensional space $\mathbf{R}^m$, and be a separating hyperplane in this space given by:

$$(\phi(\mathbf{x}_i)\cdot \mathbf{w}) + b = 0 \quad (2)$$

We want to find the minimum of:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}\xi_i \quad (3)$$

with respect to $\mathbf{w}, b$ and subject to:

$$y_i\{(\phi(\mathbf{x}_i)\cdot \mathbf{w}) + b\} - 1 + \xi_i \geq 0, \quad i = 1,\cdots,l; \quad (4)$$

$$\xi_i \geq 0, \quad i = 1,\cdots,l; \quad (5)$$

where ξ_i are the losses, $\|\mathbf{w}\|^2$ is the inverse of the class separation (margin), C represents a trade-off between the margin and the losses, and $(\cdot)$ represents a dot product in $\mathbf{R}^m$. By using the Lagrange theorem, Eq. (3) can be rewritten into:

$$L_P = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}\mu_i\xi_i - \sum_{i=1}^{l}\alpha_i(y_i\{(\phi(\mathbf{x}_i)\cdot \mathbf{w}) + b\} - 1 + \xi_i), \quad (6)$$

which has to be minimized with respect to $\mathbf{w}, b, \xi_i$, and maximized with respect to α_i, μ_i, subject to:

Departamento de Teoría de la Señal y Comunicación, Universidad de Alcalá. Ciudad Universitaria, 28871-Alcalá de Henares (Madrid, Spain).
E-mail to: antonio@ieee.org; joseluis.rojo@uah.es

$$\alpha_i, \mu_i \geq 0, \quad i = 1, \cdots, l. \tag{7}$$

The solution is a linear combination of the training data. The samples with $\alpha_i \neq 0$ are called the *Support Vectors*, and the classification function is built as a function of them.

Non-linear classifiers are built by taking the dot product in kernel generated spaces. This product uses kernels satisfying the Mercer conditions, this is, semi-defined positive kernels [3]. In this case, the problem corresponds to maximize:

$$L_d = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) + b \tag{8}$$

subject to

$$\alpha_i \geq 0; \ \sum_{i=1}^{l} \alpha_i = 0; \tag{9}$$

with classification function:

$$y = f(\mathbf{x}) = sign\left(\sum_i \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b \right) \tag{10}$$

which is known as the *Support Vector Method*. Some common kernels are:

1. Linear: $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y})$.
2. Polynomial: $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y})^d$.
3. Radial Basis Functions (RBF):

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2} \right).$$

We can see that there are two free options; the kernel parameter in non-linear classifiers (polynomial degree, *d*, or width of the RBF unit, σ), and *C* in all of them.

Vapnik [1] proposes a way of tuning these parameters from the concept Actual Risk Bound and VC dimension. For a given loss function on a classification learning problem,

$$L(y, f(\mathbf{x}, \mathbf{w})),$$

the Risk Function is defined as the mean value of this loss function:

$$E\{R(\mathbf{w})\} = \int L(y, f(\mathbf{x}, \mathbf{w})) p(\mathbf{x}, y) d\mathbf{x} dy \tag{11}$$

where $p(\mathbf{x}, y)$ is the join probability density function of the data. Vapnik shows [1] that, for a given confidence level $\eta \in [0,1]$, the following inequality is held:

$$E\{R(\mathbf{w})\} \leq R_{emp}(\mathbf{w}) + \sqrt{\frac{h(\log(2l/h)+1) - \log(\eta/4)}{l}} \tag{12}$$

where $R(\mathbf{w})$ is the *Actual Risk*, $R_{emp}(\mathbf{w})$ is the data error probability or *Empirical Risk*, and *h* is a measure of the classifier complexity, known as the Vapnik-Chervonenkis (VC) dimension of the machine. The rightmost term of Eq. 12 is known as *VC confidence*, and it represents a bound upon the complexity of the machine. The minimization of the actual risk bound has been proposed to find the optimal width or the optimal kernel parameter. However, it turns to be a weak limit in extremely low-size data sets; it sometimes needs the previous estimation of *h*; it is defined for no-losses problems, which is the least common practical case; and finally, it can not yield any information about the trade-off parameter *C*.

III. SVM Tuning Using the Bootstrap Resampling

A comprehensive formulation of the bootstrap resampling for standard error and bias estimation can be found in [4]. It is widely used today in evaluating the accuracy of statistical deals, like analysis of variance, regression models, and recently, Neural Network schemes [5]. However, it has not been taken enough advantage in tuning learning schemes, this is, in incorporating to the building process the knowledge of the nature of the estimator. This is what we are proposing here.

Be a process of estimation of dependence between the paired data of a classification problem. The data are drawn from a join fdp $p(\mathbf{x}, y)$, which is denoted:

$$p(\mathbf{x}, y) \rightarrow \mathbf{V} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_l, y_l)\}. \tag{13}$$

Be the estimated parameters through the SVM:

$$\hat{\alpha} = s(\mathbf{V}, \theta), \tag{14}$$

where $s(\mathbf{V}, \theta)$ is the process of estimation of the SVM expansion $\hat{\alpha}$. For the complete data set, the error probability is estimated through the proportion of training errors for the expansion,

$$\hat{R}_{emp} = t(\hat{\alpha}, \mathbf{V}). \tag{15}$$

A *bootstrap resample* is a set of data drawn from the training data set under their empirical distribution, this is, it corresponds to sampling with replacement the observed pairs of data:

$$\hat{p}(\mathbf{x}, y) \rightarrow \mathbf{V}^* = \{(\mathbf{x}_1^*, y_1^*), (\mathbf{x}_2^*, y_2^*), \cdots, (\mathbf{x}_l^*, y_l^*)\}. \tag{16}$$

Therefore, $\mathbf{V}^*$ will consist on elements of $\mathbf{V}$ appearing one, several or none times. The process of resampling is repeated for $b = 1, ..., B$ times. Let us consider a partition of $\mathbf{V}$ in terms of the resample $\mathbf{V}^*(b)$, such as

$$\mathbf{V} = (\mathbf{V}_{in}^*(b), \mathbf{V}_{out}^*(b)), \tag{17}$$

with $\mathbf{V}_{in}^*(b)$ representing the subset of pairs included in the resample *b*, and $\mathbf{V}_{out}^*(b)$ being the subset of non-included pairs. The obtained SVM for each resample will be given by:

$$\hat{\alpha}^*(b) = s\left(\mathbf{V}_{in}^*(b), \theta\right). \quad (18)$$

The estimation of the final parameter on this population is known as *bootstrap replication* of the statistic:

$$\hat{R}_{emp}{}^*(b) = t\left(\hat{\alpha}(b)^*, \mathbf{V}_{in}{}^*\right). \quad (19)$$

This represents an estimate of the empirical risk distribution through the B resamples. However, further advantage can be taken from the actual risk estimator, obtained by:

$$\hat{R}^*(b) = t\left(\hat{\alpha}(b)^*, \mathbf{V}_{out}{}^*\right). \quad (20)$$

The sample distribution of the replications for this last statistic represents an approximation to the true distribution of the estimated actual risk. The non-biased mean value will be obtained by simply taking the replication average. This average can be obtained for a set of values of the free parameters.

IV. Simulation Results

We generated $\mathbf{R}^{11}$ vectors, $\mathbf{v}$, being the sum of two time-varying waveforms, a *slow* plus a *fast* half-rectified, convex parabola, between 0 and 10 seconds and sampled at fs=1Hz (Figure 1), as given by the formula:

$$v(t) = k_s(t - t_s)^2 + v_s + k_f(t - t_f)^2 + v_f \quad (21)$$

$$\mathbf{v} = v(t)_{|t=0,1,\ldots,10} \quad (22)$$

Table 1 shows the rules for the random generation of the centres and the t-axis interceptions. According to the area of the slow parabola (A_1) being minor or greater than a threshold, we assigned to every vector one of two classes. The threshold was set to 3. Figure 2 depicts the probability density function of A_1. Also, a 3% of the training vectors were randomly changed their correct labels. We generated 200 training vectors. For 500 resamples the average actual risk, plus standard deviation, were estimated and compared to the error upon 10.000 test vectors. The process was repeated for a rank of possible values for each free parameter.

Figure 3 shows the performance in mean error probability (bias-corrected), standard error and number of Support Vectors for the selection of C in the linear kernel (Figure 3.a), the selection of σ in the RBF kernel (Figures 3.b and 3.c) and the selection of the polynomial degree. Some remarkable facts are:

1. There is a high coincidence between the shapes of the bootstrap estimated errors and the test errors for all the kernels.

2. For the linear and RBF kernels, when a doubt on a range of parameter values exists, the number of Support Vectors is an appropriate second criterion (Figures 3.b and 3.c).

3. The best degree for the polynomial kernel is 1, this is, the linear kernel. However, a non-linear classifier with RBF kernel yields a lower error probability. Then, this will be the best kernel among the proposed ones.

Then, the choice of the kernel parameters and C basing on the bootstrap estimated actual risk closely agree with the results on the test set.

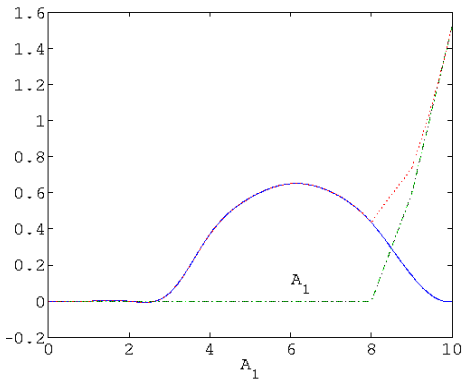

Figure 1. Parabola toy example. *Each vector (dotted) is the sampled sum of 2 parabolas, a slow (continuous) and a fast (dashed) component.*

	x_{centre}	y_{centre}	x_{in}	y_{in}
Slow	U[2,8]	U[0,1]	U[1,7]	0
Fast	10	U[1,3]	U[6,9]	0

Table 1. Parabola toy example. *Distribution of the centres and abscisa interception of the two parabolic components*

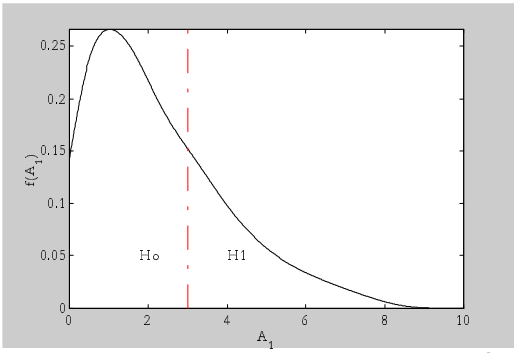

Figure 2. Parabola toy example. *Density function of A_1 and class assignment.*

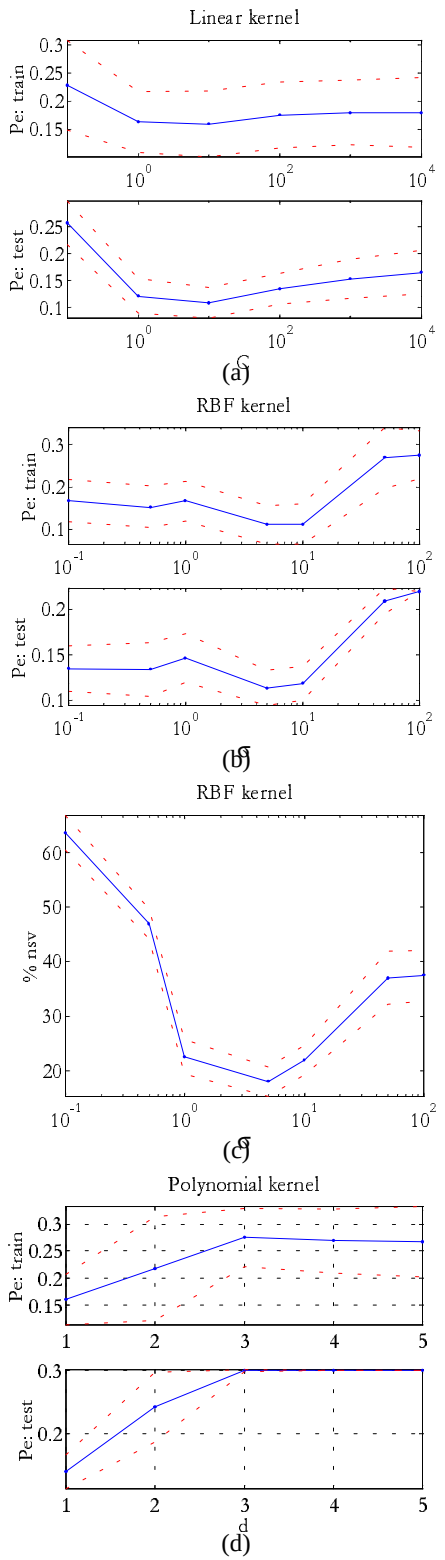

Figure 3. Parabola toy example. Results. *Mean actual risk and standard error for: (a) Linear kernel. Estimated error (up) and test set error (down) as a function of C. (b) RBF kernel. Estimated error (up) and test set error (down) as a function of σ .(c) RBF kernel. Estimated mean number of Support Vectors. (d) Polinomial kernel. The same as (b) for the polinomial degree.*

V. A Real Data Example Analysis

In this section we will perform a statistical analysis on a real set of data, comparing the use of VC-theory to the proposed bootstrap SVM tuning in a clinical problem, where the low size of the training data sets is a very common issue, due to either the cost of performing the measures or the non-presence of measure devices during the event.

We begin drawing a short introduction to the problem: the Implantable Cardioverter Defibrillator (ICD) is a device for automatic recognition and treatment of malign cardiac arrhythmia in patients with high risk of sudden death [6,7]. Present discrimination algorithms in these devices lead to a high rate of malign arrhythmia detection to the expense of a number of unnecessary delivered therapy (electric shock or cardiac stimulation) at non-malign arrhythmia episodes. So, though their safety, the discrimination algorithms in ICD are still a current framework.

The *Initial Ventricular Activation Criterion* is a recently proposed idea [8], which states the analysis of the initial voltage changes during the ventricular depolarization; according to the nature of the electric circuit involved in each type of arrhythmia, these changes will be different in malign and non-malign arrhythmia. This criterion is still been studied.

One of the issues in the statistical analysis of the changes in the ventricular depolarization is the choice of the best electrode configuration for its relative measures. Four available sources in some models of ICD are depicted in Figure 4:

1. HVA/HVB: between the subpectoral can and the defibrillation coil;
2. HVA/P+S: between the subpectoral can and the sensing distant dipole;
3. P-S/HVB: between the sensing proximal dipole and the defibrillation coil; and
4. P-S/P+S: between the proximal and distant dipole.

A set of morphological parameters have been proposed for featuring these changes in the electrical activity. These parameters consisted on amplitudes, time activations and energies on the first derivative of the initial electrical ventricular depolarization waveform. The parameters were used for the purpose of classification between the tachycardia (T) episode and the preceding Sinus Rhythm (SR) or normal cardiac depolarization. We tested the distance between these two rhythms for each source through these parameters.

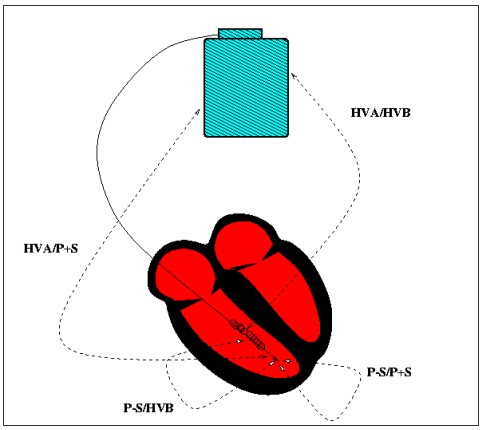

Figure 4. Electrode configuration selection. *Scheme of the available electrode configuration in the ICD.*

On a number of tachycardia episodes, the parameters were measured from the 4 available electrode sources, when possible. This led to 56 records for HVA/HVB, 52 for HVA/P+S, 54 for P-S/HVB and 54 for P-S/P+S. A classifier was built in order to discriminate the parameter vectors measured during SR from those ones measured during T. We compared the next schemes:

1. The Vapnik's bound, as given by Equation (13), on a linear kernel SVM classifier. For this purpose, the empirical risk was calculated as the error ratio in the training set for each one of the classifiers. The structural risk (VC confidence) was estimated through the approximation of the VC-dimension by the number of Support Vectors found by the linear classifier. The same value of C=10 was set for all the classifiers, considered heuristically appropriate. This corresponded to the same trade-off between margin and losses among all the classifiers, given that the number of samples in each source is approximately the same.

2. The bootstrap-estimated error probability on non-linear kernel (RBF) SVM classifiers. For this case, the gaussian width was previously determined on every classifier through 20 resamples of the training data set for a range of values in (0.1,100). The value of C was also previously determined through 20 resamples on every classifier, for a range of values in (0.1,500). Both ranges were swept through a logarithmic scale. The two free parameters were considered jointly, as far as influence between them had been observed. The best pair of values was found recursively for each source case.

The parameters were normalized by their standard deviation, in order to avoid the effect of large scale features.

Source	$R_{emp}(\alpha)$	*VCcon*	$R_{act}(\alpha)$	*RBF*
HVA/P+S	0.032	0.590	0.622	0.008
P-S/HVB	0.053	0.614	0.668	0.054
P-S/P+S	0.125	0.632	0.757	0.125
P-S/P+S	0.231	0.623	0.854	0.102

Table 2. Electrode configuration selection. Empirical risk, VC confidence and Vapnik's bound on the error probability for a SVM with linear kernel; empirical error probability for the SVM RBF kernel

.

Table 2 shows the results for both schemes. As it can be seen there:

- The error probabilities for the linear SVM are sorted in growing order. This points to the fact that HVA/HVB is the most far-field character electrode configuration, P-S/P+S is the most near-field configuration, and the other are intermediate between them. So, the linear analysis points to far-field sources as the most appropriate for this criterion.

- However, the VC confidence takes values upon 0.5 in all the cases. Strictly speaking, these classifiers are worst than a coin-based choice. Nevertheless, the qualitative result is according to the empirical risk on the linear kernel.

- When the RBF kernel is used, the bootstrap error yield to more realistic estimations for the actual risk. More, there is a reduced error in the P-S/P+S configuration, which points to a possible influence of the kernel in the error probabilities.

Consequently, neither the empirical risk nor the Vapnkik's actual risk bound represent adequate statistical error measures, despite of their qualitative coincidence. Bootstrap measures on SVM allow to perform non-linear analysis on low-sized data sets.

VI. CONCLUSSIONS

We conclude that this understanding of the use of the bootstrap resampling leads to an accurate estimation of the risk on a learning machine, and to a solid way of estimating the free parameters on a SVM based learning machine.

VII. BIBLIOGRAPHY

[1] Vapnik, V. Statistical Learning Theory (Adaptive and Learning Systems for Signal Processing, Communications and Control). John Willey&Sons, 1998.

[2] Schölkopf, B., Burges, C., Smola, A., editors. Advances in Kernel Methods-Support Vector Learning. MIT-Press. 1999.

[3] Burges, C. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2(2):1-32. 1998.

[4] Efron, B., and Tibshirani, R. An Introduction to the Bootstrap. Chapman&Hall, 1993.

[5] Tibshirani, R. A Comparison of Some Error Estimates for Neural Network Models. Technical Report, Dept. of Preventive Medicine and Biostatistics, University of Toronto, Toronto, Ontario, 1994.

[6] Singer, I. Implantable Cardioverter Defibrillator. Futura Publishing Co. Inc., 1994.

[7] Smith, W.M., and Ideker, R.E. Automatic Implantable Cardioverter Debrillators. Annual Rev. Biomed. Eng., 1(1):331-46, 1999.

[8] Rojo-Álvarez, J.L., Arenal, A., Artés, A., et al. Discrimination Between Ventricular and Supraventricular Tachycardia Based on Implantable Defibrillator Stored electrogram Análisis. In 47 American College of Cardiollogy, pag. 294-A, Atlanta, march 1998.